Reverse-engineering censorship in China: Randomized experimentation and participant observation

Reverse-engineering Censorship in China: Randomized Experimentation and Participant Observation*

Gary King† Jennifer Pan‡ Margaret E. Roberts§

This is the draft version of our paper; the final published version of this paper is available at http://j.mp/ChinaExp

*For helpful advice, we thank Peter Bol, Sheena Chestnut, Peter Gries, Yoi Herrera, Haifeng Huang, Iain Johnston, Susan Shirk, Dustin Tingley, and participants in a panel at the American Political Science Association, August 31, 2013 and at the Midwet Political Science Association April 3, 2014. For expert research assistance over many months, we are tremendously appreciative of the efforts and insights of Frances Chen, Wanxin Cheng, Amy Jiang, Adam Jin, Fei Meng, Cuiqin Li, Heather Liu, Jennifer Sun, Hannah Waight, Alice Xiang, LuShuang Xu, Min Yu, and a large number of others who we shall leave anonymous. We thank Crimson Hexagon, Inc. for help with data. Replication data and information is available in King, Pan, and Roberts (2014).

†Albert J. Weatherhead III University Professor, Institute for Quantitative Social Science, 1737 Cambridge Street, Harvard University, Cambridge MA 02138; http://GKing.harvard.edu, king@harvard.edu, (617) 500-7570.

‡Ph.D. Candidate, Department of Government, 1737 Cambridge Street, Harvard University, Cambridge MA 02138; http://people.fas.harvard.edu/∼jjpan/, (917) 740-5726.

§Ph.D. Candidate, Department of Government, 1737 Cambridge Street, Harvard University, Cambridge MA 02138; http://scholar.harvard.edu/mroberts/home

Abstract

Existing research on the extensive Chinese censorship organization uses observational methods with well-known limitations. We conduct the first large-scale experimental study of censorship by creating accounts on numerous social media sites, randomly submitting different texts, and detecting from a worldwide network of computers which are censored. We also supplement the usual interviews with secret sources by creating our own social media site, contracting with Chinese firms to install the same censoring technologies as existing sites, and — with their software, documentation, and even customer support — reverse engineering how it all works. Our results offer rigorous support for the recent hypothesis that criticism of the state, its leaders, and their policies are published, whereas posts about real world events with collective action potential are censored.

1 Introduction

The Chinese government has implemented "the most elaborate system for internet content control in the world" [1], marshaling hundreds of thousands of people to strategically slow the flow of certain types of information among the Chinese people. Yet, the sheer size and influence of this organization has made it possible for researchers to infer via passive observation a great deal about its purpose and procedures, as well as the intentions of the Chinese government. We seek to get around the well known inferential limitations inherent in observational work by large scale random experimentation and participant observation.

We begin here with the theoretical context. The largest previous study of the purpose of Chinese censorship distinguished between the "state critique" and "collective action potential" theories of censorship and found that, with few exceptions, the first was wrong and the second was right: unlike most prior claims, even vitriolic criticisms of the government in social media are not censored but any attempt to physically move people in ways not sanctioned by the government are. Even posts that praise the government are censored if they pertain to real world collective action events [2].

In both theories, regime stability is the assumed ultimate goal [3, 4, 5, 6]. For example, scholars had previously thought that the censors pruned the Internet of government criticism and biased the remaining news in favor of the government, thinking that others would be less moved to action on the ground as a result [7, 8, 9]. However, even if biasing news positively would in fact reduce the potential for collective action, this state critique theory of censorship misses the value to the central government and central Party organization of the information content provided by open criticism in social media [10, **?**, **?**, 11]. After all, much of the job of leaders in an autocratic system is to keep the people sufficiently mollified so they will not take action that may impact their hold on power. In line with the literature on responsive authoritarianism, knowing that a local leader or government bureaucrat is engendering severe criticism, perhaps because of corruption or incompetence, is valuable information [**?**, **?**]. That leader can then be replaced with someone more effective at maintaining stability, and the system can then be seen as

responsive. This responsiveness would seem likely to have a considerably larger effect on reducing the probability of collective action than merely biasing the news in predictable ways.

The collective action potential hypothesis holds that the Chinese censorship organization first detects a volume burst of social media posts within a specific topic area, and identifies the real world event that gives rise to the volume burst [2]. If the event was classified as having collective action potential, then all posts within the burst were censored, regardless of whether they were critical or supportive of the state and its leaders. Unlike the uncertain process involved in coherently classifying individual posts as to their collective action potential, this procedure is easily implemented with extremely high levels of inter-coder reliability. No evidence exists as to whether these rules were invented and directed by a person or committee in the Chinese government are merely an emergent pattern of this large scale activity.

Although the largest existing study analyzed study analyzed more than 11 million social media posts from almost 1,400 web sites across China [2], it along with other quantitative studies of censorship [12, 13] are solely observational, meaning that some conclusions necessarily depend upon untestable assumptions. For example, the data for these studies is controlled by an earlier stage where many social media web sites use automated review (based on techniques like keyword matching) to immediately move large numbers of prospective posts into a temporary limbo to receive extra scrutiny before possible publishing (for a guide, see Figure 1). Whereas the ex post content filtering decision is conducted largely by hand and takes up to about 24 hours, the ex ante decision of whether posts are slotted for review is automated, instantaneous, and thus cannot be detected by observational methods. Importantly, this also means that the automated review process could induce selection bias in existing studies of censorship which can only observe those submissions that are not stopped from publication by automated review. And of course observational research generally also risks endogeneity bias, confounding bias, and other problems.

To avoid these potential biases, and to study how automated review works, we conduct

a large scale experimental study, where random assignment controlled by the investigators substitutes for statistical assumptions. We do this by creating accounts on numerous social media sites across China; writing a large number of unique social media posts; randomizing assignment of different types of posts to accounts; and, to evade detection, observing from a network of computers all over the world which types are published or censored. Throughout, we attempted to avoid disturbing the flow of normal discourse by producing social media content on topics similar to those in real social media posts (including the content of those censored, which our methods can access). Although very small scale nonrandomized efforts to post on Chinese websites and observe censorship have been informative [14], randomized experiments have not before been used in the study of Chinese censorship.

In addition to our randomized experiment, which we use to make causal inferences, we also seek to produce more reliable descriptive knowledge of how the censorship process works — information important in its own right, intensely studied and contested in the academic and policy communities, and ultimately essential also for any causal study. Gathering this information, until now, has mostly come from highly confidential interviews with censors or their agents at social media sites or in government, information that is necessarily partial, incomplete, potentially unsafe for research subjects, and otherwise difficult to gather. We add a new source of information that has not been attempted before in studies of censorship through participant observation. Most importantly, this enabled us to change the incentive structure of our informants. Thus, from inside China, we created our own social media website, purchased a URL, rented server space, contracted with one of the most popular software platforms in China used to create these sites, submitted, automatically reviewed, posted, and censored our own submissions. The website we created is not available to anyone other than our research team to avoid affecting the object of our study or otherwise interfering with existing Chinese social media discourse. However, in doing so, we had complete access to the software, documentation, help forums, and extensive consultation with support staff; we were even able to get their recommendations on how to conduct censorship on our own site to adhere to government standards. The

"interviews" we conducted in this way were unusually informative because the job of our sources was in fact to answer the questions we posed.

Overall, this work offers three intended contributions. First, by analyzing large numbers of posts at numerous social media sites, we are able to resolve some disagreements in the policy and academic literatures on the subject, such as explanations for the presence of conflicting keyword lists and the absence of a coherent or unified interpretation for the operation of these lists at individual sites. Consistent with this disagreement, we show that the large number of local social media sites in China have considerable flexibility, and choose diverse technical and software options, in implementing censorship. Second, we show that the automated review process affects large numbers of posts on fully two-thirds of Chinese social media sites, but is a largely ineffective step in implementing the government's censorship goals. This is surprising but consistent with the known poor performance of most keyword-based approaches to text classification. Finally, despite automated review's large presence, high potential for generating selection bias in observational studies, and overall ineffectiveness due to keyword matching, we find that the government is still able accomplish its objectives — as summarized by the collective action potential hypothesis — by using very large numbers of human coders to produce post hoc corrections to automated review and to censorship in general. Our research offers clear support for the collective action potential hypothesis and then offers some significant extensions. We find, consistent with the implications of this theory, but untested in prior research, that posts about collective action events outside mainland China, collective action events occurring solely online, social media posts containing critiques of top leaders, and posts about highly sensitive topics such as Tibet and Xinjing not during collective action events, are not censored.

In Section 2, we summarize our experimental designs, and the unusual logistical difficulties in engineering and executing them in this context (with additional details in the supplementary materials). This design section also covers our participant observation approach in creating a social media site, which helps us more carefully define the process we will experiment on. Section 3 presents our results and Section 4 pushes the collective ac-

tion potential theory until it breaks so that we can find the edges of where it is applicable. Section 5 concludes.

2 Research Designs

We now describe the challenges involved in large scale experimentation, participant observation, and data collection in a system designed to prevent the free flow of information, especially about the censors. These include avoiding detection so we were not prevented from carrying out our study, implementation on the ground in many geographically distant places, keeping a large research team safe, and ensuring that we do not disturb or alter the system we are studying. The human subjects aspects of our experimental protocol were pre-approved by our university's Institutional Review Board. For obvious reasons, we are unable to reveal certain details of how we implemented this design, but we do give complete information on the statistical and scientific logic behind our choices, which are straightforward.[1]

We begin with the outcome variable we are studying and then describe our experimental protocols.

2.1 Participant Observation

Aspects of the process by which censors in the Chinese government and social media companies implement censorship directives have been gleaned over the years in interviews with sources that have first hand knowledge, including the censors themselves. We have also conducted many such interviews, and each one produces some information, but it is necessarily a partial picture, highly uncertain, and potentially unsafe for the sources and researchers.

Thus, we looked for a way to learn more by changing the incentives of our sources. We did this by creating our own Chinese social media site from inside China, using all the

[1] We also added our own ethics rules, not required by the IRB, which dictates that we avoid, wherever possible, influencing or disturbing the system we are studying. The similarity to The Prime Directive in Star Trek notwithstanding, this seems like the appropriate stance for scientists attempting to understand the world, as distinct from advocates trying to change it, and in any event is more likely to yield reliable inferences. Further details can be found in our supplementary materials.

infrastructure, procedures, and rules that existing sites must follow. To do this, we purchased a URL, contracted with a company that provides hosting services, and arranged with another company to acquire the software necessary to establish a community discussion forum (a Bulletin Board System (BBS)). We downloaded the software and installed it ourselves. This infrastructure gave us complete access to the software and its documentation so that we could fully understand and utilize its functionality. Importantly, we also had easy access to support employees at these firms, who were happy to help show us how to censor so that our website remained in accordance with their view of government requirements. Thus, instead of trying to convince people to spare some of their time for researchers, we were able to have conversations with employees whose jobs it is to answer questions like those we posed, and fortunately they seem quite good at their jobs. We then studied and customized the software, submitted posts ourselves, and used the software's mechanisms to censor some. We took every step we could short of letting individuals in China post on the site to avoid causing any interference to actual social media discourse.

The biggest surprise we found relative to the literature was the huge variety of technical methods by which automated review and human censorship can be conducted. Table 1 summarizes some of these options.

When we installed the software, we found that, by default, it included no automated review or blocking. But webmasters can easily change the option of automatically reviewing specific types of users (those who are moderators, super users, users who have been banned from posting, or those who have been banned from visiting the site), internet protocol (IP) address, new threads, or every response — all of which can be tailored for each of as many forums as is set up on each website. Functionality also exists to bulk delete posts, which can be implemented by date range, user name, user IP, content containing certain keywords, or by length of post. On the backend, the webmaster also has flexible search tools to examine content, to search by user name, post titles, or post content. What the user sees can also be limited: the search function can be disabled for users, webmasters have the option of whether to allow users to see whether or what posts of theirs are being automatically reviewed.

Table 1: Options for Content Filtering on Forum Platform

1. Automated Review Options	
Content-based review can be based on:	
- moderator-supplied keywords	- review specific to post type (e.g. comment or main post)
-plugins for reviewing posts with minimal influence on the user	- review specific to forum topic
-plugins advertising better keyword blocking technology	
User-based review can be based on:	
- user IP	- previous user posts
- payments by user	- last login
- points won by user (e.g. for number of posts, comments)	
Time-period review and censorship allows:	
- periods of time where all posts are audited	- disallow posting during certain hours of the day
Workflow for reviewed posts:	
- different censors for different types of postings (e.g. spam vs. political content)	- review interface with search functionality
- batch deletion of posts	
2. Account Blocking Options	
- blocking for specific types of posts (e.g. comment or main post)	- blocking based on user IP
-blocking for specific forums	- blocking posting and/or reading
- blocking based on points	

We found employees of the software application company to be forthcoming when we asked for recommendations as to which technologies have been most useful to their other clients in following government information management guidelines. Based on their recommendations, as well as user guides, detailed analyses from probing the system, and additional personal interviews (with sources granted anonymity), we deduce that most social media websites that conduct automatic review do so via a version of keyword

matching, probably using hand-curated sets of keywords (we reverse engineer the specific keywords below).[2]

Based on what we learned, we summarize the censorship process in Figure 1. The process begins when one writes and submits a blog or microblog post at a social media web site (left). This post is either published immediately (top left node) or held for review before publication (middle left node in red). If the post is published immediately, it may be manually read by a censor within about 24 hours and, depending on the decision, either remains online indefinitely (top box) or is removed from the Internet (second box). As can be seen from the screen shots of actual web sites in Figure 1 (with full examples in the supplementary materials), the decisions of the censors, and the fact that they are by the censors, are unambiguous.

The censors then read each post in review (usually within a day or two) and either publish the post (third box of Figure 1) or delete it before publication (fourth box of Figure 1). We are able to identify review when it occurs because typically a message is given after post submission that informs the user the text has been slotted for review. In the absence of a warning message, the user can tell when a post is put into review because no public URL is associated with the post, and the user's account page will show the status of the post as "under review". Finally, on the basis of the current and previous posts, a submitted post can be censored and the account blocked so that no additional posts may be made (last box of Figure 1). In this last case, when a user submits a text for posting, an error message notifying the user of account blocking is encountered. A key point is that the massive data set in [2] corresponds only to the first three boxes, whereas in our experiment we are able to study all five paths down the decision tree.

2.2 Experimental Protocol

We now give the experimental protocol which we designed to make causal inferences without certain modeling assumptions. We first selected 100 social media sites, including 97 of the top blogging sites in the country, representing 87% of blog posts now on

[2]In the process of setting up the site, they recommended that we hire 2-3 censors for every 50,000 users. That enables us to back out an estimate of the total number of censors hired within firms at between 50,000 and 75,000, not counting censors within government, 50 cent party members, or the Internet police.

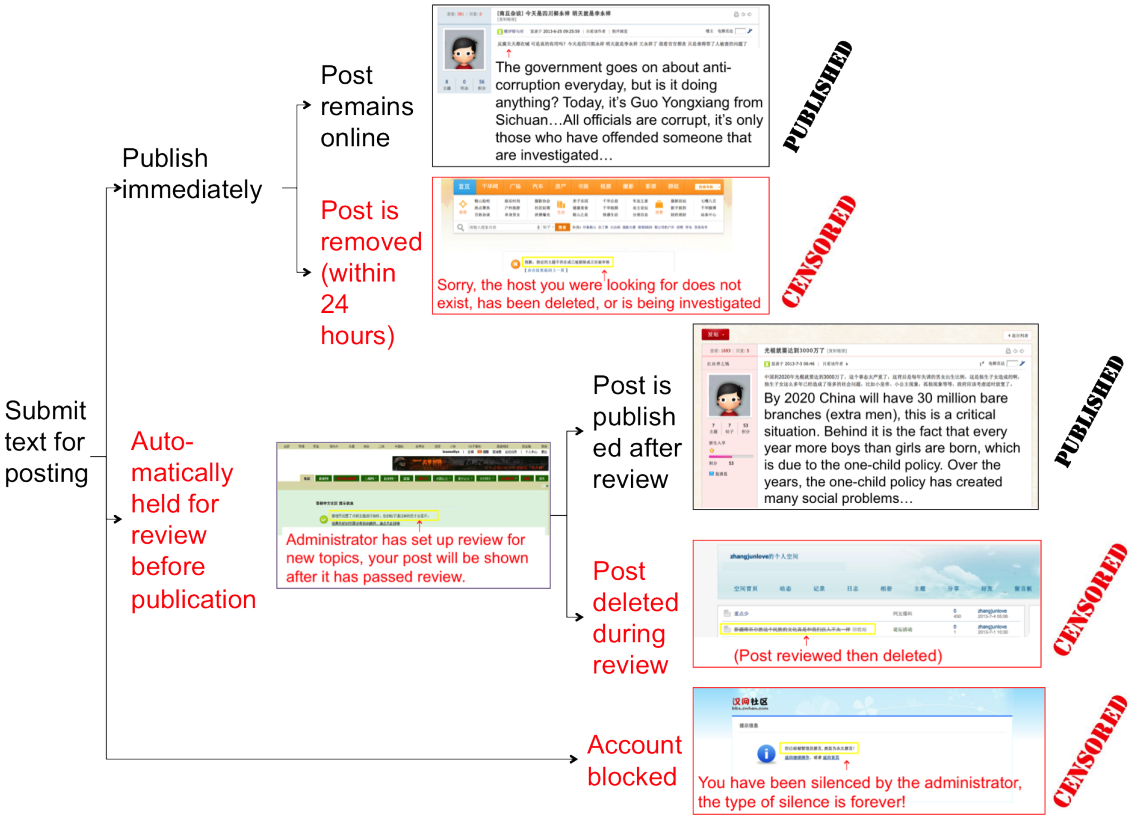

Figure 1: The Chinese Censorship Decision Tree: The pictures shown are examples of real (and typical) web sites, along with our translations. Observational studies are based only on the first three paths through this decision tree; our experimental study includes all five. Full screen shots are in our supplementary materials.

the web, and including the top three microblogging (i.e., Twitter-like) sites, sina weibo (weibo.com), tencent weibo (t.qq.com), and sohu weibo (t.sohu.com). The first two of these microblogging sites each include over 500 million registered users, and 50-100 million daily active users.[3] Together, the 100 sites are geographically spread all over China and are divided among those run by the government ($n_g = 20$), state-owned enterprises ($n_s = 25$), and private firms ($n_p = 55$). Some cater to national audiences, whereas some only allow those local to post. Creating accounts on some of these sites requires the user to be in the country, at a specific geographic locale, have a local email address, or provide another method of communication for identification. We devised procedures to create two accounts at each of these social media sites.

We kept our design close to aspects of [2]. The theory in that paper was not that every

[3]See [15, 16] for numbers of registered users which are substantial even if we account for automated sites created by marketing firms [17].

social media post with the potential to generate collective action is censored. After all, almost any issue could in principle be used as a hook to generate protest activity. Instead, the theory is that (pro- or anti-government) posts concerning a collective action event are censored. Collective action events are those "which (a) involve protest or organized crowd formation outside the Internet; (b) relate to individuals who have organized or incited collective action on the ground in the past; or (c) relate to nationalism or nationalist sentiment that have incited protest or collective action in the past." [2, p.6].

We conducted three rounds of experiments (April 18-28, June 24-29, and June 30-July 4, 2013), during which social media posts would need to be written in real time about current issues, making the logistics challenging. At the beginning of each round, we scoured the news and selected on-going collective action events and non-collective action events about which there was a significant volume of social media discussion. We chose a ratio of one collective action event to two non-collective events, since collective action events are more scarce and so that we can average over different non-collective action events. We included non-collective action events only if they were widely discussed topics pertaining to actions taken by the Chinese government, officials, or the Chinese Communist Party (CCP), which are unrelated to events with collective action potential. We also attempted where possible to select events that mentioned specific officials' names and addressed what has been described as especially "sensitive" topics. (We also included several edge cases described in Section 4.) Details of all events appear in the supplementary materials, but here are the four collective action events we found when our study was conducted, all of which meet the definition but some of which are more incendiary than others:

1. Qui Cuo, a 20 year old mother self-immolated to protest China's repressive policies over Tibet. Her funeral drew protesters.

2. Protesters in Panxu, a village in Xiamen Fujian, took to the streets because they claim officials did not adequately compensate them for requisitioning their collectively owned farmland to build a golf course. Village representatives went to local authorities to demand compensation but were instead detained. Thousands of villagers went to the town hall to demand the release of the village representatives,

police moved in to arrest the villagers, villagers retaliated by smashing police cars, and taking the local Party secretary into custody.

3. On the second anniversary of the 2011 arrest of artist-dissident Ai Weiwei, he released an album that talks about his imprisonment. Ai Weiwei was arrested in 2011 on charges of tax evasion, but more likely either for calling his followers to mimic the Arab Spring or for organizing volunteers to collect the names of children who died in the Sichuan earthquake. The release of the album by Ai Weiwei is chosen as an example of collective action under part (b) of the definition, where posts about individuals who have organized or incited collective action on the ground in the past are censored.

4. An altercation between Uyghurs (a minority ethnic group) protesting and local police in Lekeqin township of Shanshan county in Turpan, Xinjiang. 24 were killed, including 16 Uyghurs. Police and many official news reports of the event attribute it as an act of Uyghur terrorism, but rumors circulated in social media that the protest was precipitated by forced housing demolition.

For each event, we had a group of native Chinese speakers write some posts supportive and others critical of the government based on example social media posts that had already appeared online, including posts which were censored as well as those that remained online. (We used the technology of [2] to obtain access to the censored posts.) In other words, we obtain posts that are immediately published after submission, including those which remain online and those which are removed (top two boxes of Figure 1). We provided our writers with background on the event, the definition of what we mean by pro- and anti-government for each topic (see the online supplement), and examples of real posts from Chinese social media similar to those we needed written. So that we could minimize any experimenter effect, we checked each text ourselves by hand and attempted throughout to ensure that the posts we submitted were similar in language, sentiment, and content to those already found in (or written and censored in) Chinese social media.

From a statistical point of view, we blocked [18] on three variables: First, our posts included the same keywords in both the treatment and control conditions. Second, we

controlled for individual writing style by blocking on author in our experimental design. That is, posts in each set of four experimental conditions (defined by our two variables: pro/anti government, and with/without collective action potential) was authored by the same set of research assistants. And finally, we constrained all posts to be between 100 and 200 characters in length. In addition, we also ensured that no two posts submitted were exactly identical to each other or to any we found in social media. All posts were submitted between 8am and 8pm China time from the U.S. or from the appropriate place within China, depending on what was feasible because of the technology used at each social media site.[4]

We were interested in testing the causal effect of both pro- vs. anti-government content and collective action vs. non-collective action content, leading by cross-classification to four logical treatment categories. To make the most efficient use of each individual account, we submitted two posts to each. But it makes little sense for one account (representing a single person) to write both pro- and anti-government posts regarding the *same event*. Thus, we submitted posts about two different events to each account which were pro-government collective action and anti-government noncollective action, or instead anti-government collective action and pro-government non-collective action. In this way, every account contributes to the causal effect estimate of each hypothesis. We also ensured our ability to make causal inferences without extra modeling assumptions by randomizing (a) the choice between these two pairs, (b) the order within each pair, and (c) the specific collective action and policy events we wrote about in each submission. Missingness can occur when web sites are down, if an account we created expired, or if an account is blocked due to prior posts. Largely because of the design, any missingness will be almost exactly independent of our two treatment variables; empirically that proved to be the case.

Each of the 100 different social media web sites in our study offers different ways of

[4]All posts were made to mainland China accounts. Some were submitted from outside China, when it feasible, and many from within China. Recent work has noted that overseas accounts are subject to less stringent censorship regulations than mainland accounts [19]. This issue does not affect our work since all accounts created and used are mainland China accounts. Users can control account location when creating the account by specifying a location in China, by entering local mobile number, or by creating the account from a local IP address. We used all of the latter methods.

expressing oneself online. When possible, we submit posts on the home page we created for each account. For discussion forums, we start a new thread with the content of the post in the most popular sub-forum. On sites where creating new threads by users is not permitted, we submit posts as a reply to an existing thread relevant to the topic. In all cases, we write our posts so as not to stand out from the stream of existing information, following all social media, web site, and cultural norms. In total, we wrote 1,200 posts by hand, every one unique, and none referring to each other.[5]

After submitting a post, we observed whether it was caught by the process of automated review; if in automated review whether and when it was eventually published; and if not caught by the automated review process whether it was eventually censored after the fact or it remained on the web. When a post appeared on the web, we recorded the URL and verified censorship from computers inside and outside of China. We recorded the outcome in terms of censorship, which corresponds to the branches of the decision tree in Figure 1.

Throughout, our goal was that anyone looking at the submissions we wrote would have no any idea this was part of an academic research project, was not different than what they might find otherwise, and would not in any way disrupt or change the social media ecosystem we were studying. We also needed to ensure that our checking published posts for censorship was not obtrusive. So far as we are aware, no one outside of our research team and confidants were aware of this experiment before we made this paper available, and no one on the web indicated any suspicion about or undue attention toward any of our posts.

3 Results

We find that in aggregate, automated review affects a remarkably large portion of the social media landscape in China. In total, 66 of the 100 sites in our sample (automatically)

[5]For each of our three rounds, we wrote 200 posts on non-collective action events (split equally between pro- and anti-government), 200 posts on collective action events or edge cases described in Section 4 (again split equally between pro- and anti-government). Thus, 600 posts submitted relate to non-collective action events, and 600 relate to collective action events or edge cases. We have in total 4 collective events and 2 edge cases, and so 400 posts focused on collective action events, and 200 on edge cases.

review at least some social media submissions, and 40% of *all* of our individual social media submissions from our 100 sites (and 52% of submissions from sites which review at least sometimes) are put into review. Of those submissions which go into review, 63% never appear on the web.

These figures indicate that automated review affects a large component of intended speech in China and so deserves systematic attention from researchers. This is especially so because of conflicting conclusions and lack of a unified interpretation in the academic and policy literatures about which keywords provoke action by the government, how automated review works, and what impact this process ultimately has on the content of speech which is blocked and which can be consumed by the Chinese people [20, 21]. We offer a possible resolution to these issues here.

3.1 Censorship

Using our broader sample, unaffected by selection during the automated review process, and with our experimental randomization, we begin by testing the collective action potential hypothesis. Based on a difference in means between the treatment and control groups, the black dots in the left panel of Figure 2 summarize the point estimate for the causal effects on censorship of submitting posts about four separate collective action events. The vertical lines are 95% confidence intervals (as with all our figures). The effects are substantial, ranging from about 20 to 40 percentage point differences (denoted on the vertical axis) solely due to writing about an ongoing collective action event as compared to an ongoing noncollective action event.

We also go a step further examine some of the other decision paths in Figure 1. To do this, we estimate the "causal mediation effect" [22, 23] of submitting posts about collective action events (vs noncollective action events) on censorship and find that almost none of this effect post content is mediated through automated review: the overall effect is a trivial 0.003 probability, with a 95% confidence interval of $(-0.007, 0.016)$; details appear in the Supplement. The (non)effect for each of the four collective action events we studied is displayed in the right panel of Figure 2, and each is similarly approximately zero, with a small confidence interval. Review, which appears to be fully automated, is

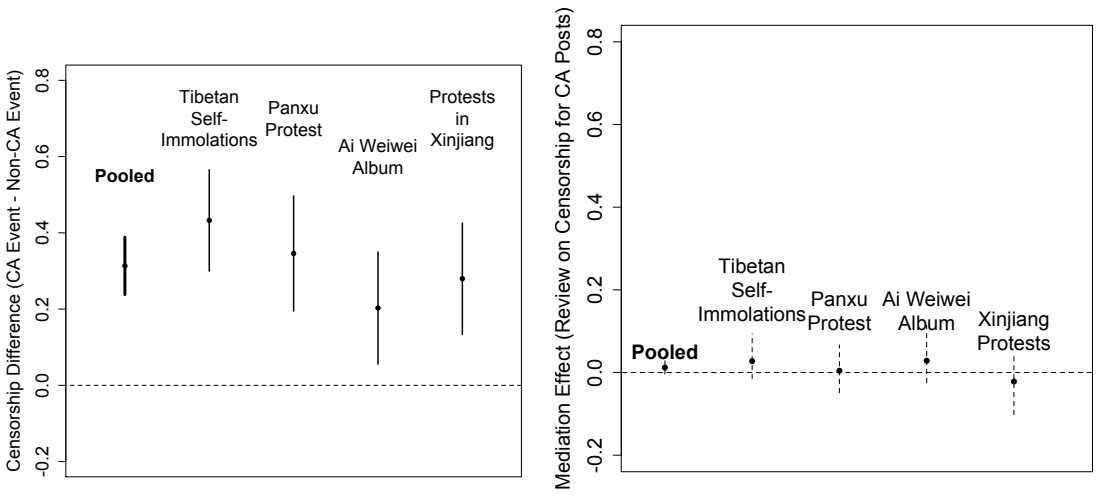

Figure 2: The Causal Effect on Censorship of Posts with Collective Action Potential (left panel) and The Mediation Effect of Review (right panel): Collective action events are more highly censored than non-collective action events within the same time period. However, censorship of collective action events is not mediated through automated review.

thus applied in a manner independent of other relevant variables and like most keyword-only methods of automated text analysis, it does not appear to work well at scale. From this result, it even appears that the censors largely ignore it or at least do not get much information from it. (We study this in more detail in the next section.)

In parallel to the large causal effect for collective action, Figure 3 reports tests of the state critique hypothesis for each of our four collective action events and eight (non-collective action) policy events. The black dots summarize point estimates of the causal effect of submitting posts in favor of the government vs opposed to the government about each event. As can be seen, the dots are all very close to the horizontal dashed line, drawn at zero effect, with six dots above and six below, and all but one of the confidence intervals crossing the zero line. Note especially that there is no hint of more censorship of anti-government posts when they involve topics that might be viewed as more sensitive or which specifically mention the names of Chinese leaders (see Supplementary Appendix for contextual details). This finding runs counter to anecdotal evidence that rumors and names of leaders unrelated to collective action lead to censorship.

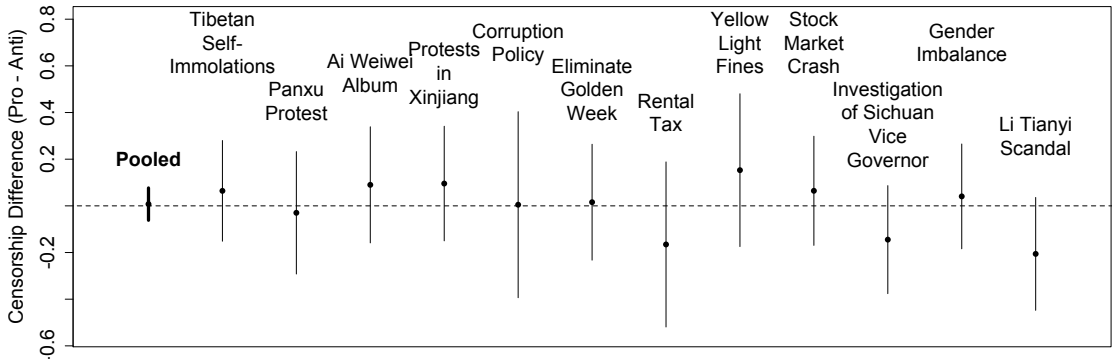

Figure 3: The Causal Effect on Censorship of Posts For or Against the Government: Posts that support the government are not more or less likely to be censored than posts that oppose the government, within the same topic.

3.2 Automated Review

The overall results in favor of the collective action potential hypothesis and against the state critique hypotheses thus appear unambiguous. The automated review process has a nearly undetectable effect on evidence about that hypothesis, since the human censors correct errors after the keyword matching techniques are applied in automated review (even though even incorrect keyword filtering slows down communications on many subjects). We now go back up the decision tree of Figure 1 to study the automated review process more directly.

We first notice that not all websites have automated review turned on, and that the method of censorship varies enormously by website (this is also true for account blocking, which we describe further in the online supplement). This is consistent with what we learned from creating our own social media site, where the software platform not only allows the option of whether to review, but also offers a large variety of choices of the criteria by which to review. Indeed, there exists considerable diversity in the technologies used by different social media sites for automated review [13]. It is this diversity in technology across sites, then, which appears to account for why different researchers typically find different patterns when looking at different sites, or at specific issues, such as which keywords are used to detect posts to be caught in automated censorship processes. This also accounts for why researchers have been unable to offer unified interpretations of what they observe consistent with reasonable assumptions about the goals of the Chinese

leadership. It is only by looking at the whole process at scale does the simplicity of the Chinese government's goals become clear.

Why would the government allow for a free choice from a large number of censorship methods, in the course of providing top down, authoritarian control? To answer this question, we collected detailed information about all software platforms and plugins available for purchase or license by social media sites to control information. From this study, we conclude that the government is (perhaps intentionally) promoting innovation and competition in the technologies of censorship. Such decentralization of policy implementation as a technique to promote innovation is common in China [24, 25, 26, 27].

Based on interviews with those involved in the process, we also find a great deal of uncertainty over the exact censorship requirements and the precise rules for which the government would interfere with the operation of social media sites, especially for smaller sites with limited government connections. This uncertainty is in part a result of encouraging innovation, but it may also in some situations be a means of control as well–it being easier to keep people away from a fuzzy line than a clearly drawn one.

We begin a systematic empirical study by understanding which social media websites use any automated review process. Figure 4 presents a density estimate (a continuous version of a histogram) of the distribution of the proportion of posts reviewed for three types of sites, depending on ownership. As can be seen, it is government sites that have the highest probability of a post being put into review, followed by the state owned enterprises, followed last by privately owned sites (which tend to have the largest user bases).

Why would government sites be more likely to delay publication until after automated review, whereas private sites publish first and make censorship decisions later? So far as we can tell from qualitative evidence, the reason is the penalty for letting offending posts through differs between government and private sites. A government worker who fails to stem collective action could lose his or her job immediately; in contrast, a worker in a private site that makes the same mistake cannot usually be directly fired by the government. Indeed, government workers have a historical legacy of prioritizing following orders and not making mistakes, even if it is considerably more inefficient to do so [28]. Private

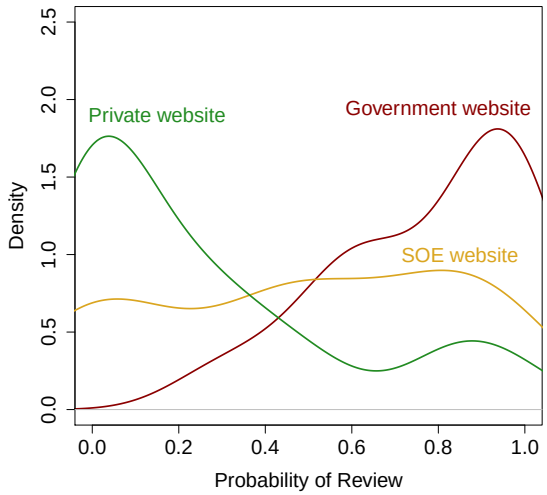

Figure 4: Histogram (density estimate) of the proportion of posts reviewed by site. The graph shows that government-controlled social media sites catch more posts by automated review much more than privately owned sites; social media sites controlled by State-Owned Enterprises (SOE) are in the middle.

sites, on the other hand, have incentives to publish as much as they can so as to attract more users. A private site can of course be taken down entirely, but that kind of "nuclear option" is used less often than more generalized pressure on the leadership of the private social media sites.

What are these largely government sites reviewing? In a manner directly parallel to Figures 2 and 3 for the ultimate variable of censorship, we now conduct an analysis of the effects on automated review of collective action and pro and anti-government posts. Figure 5 gives results for the effect of collective action on review: they include four positive estimated effects but two are small and three are have zero inside their confidence intervals. If the goal of the censors is to capture collective action events, the automated algorithm is performing marginally at best, although this is quite common for keyword algorithms which tend to work well for specific examples for which they can be designed but often have low rates of sensitivity and specificity when used for large numbers of documents.

Also interesting is the causal effect of pro- vs anti-government posts in Figure 6. These are all small, and most of the confidence intervals cross zero. In fact, if there exists a

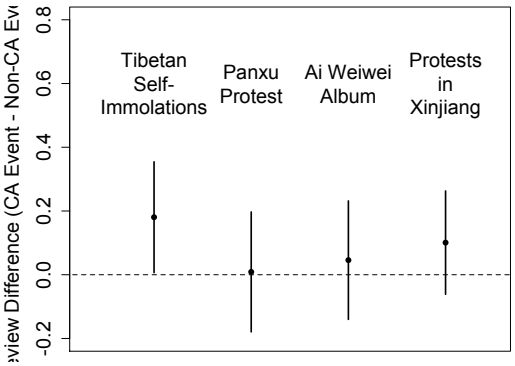

Figure 5: Causal Effect on Review of Collective Action Potential Events: Collective action events are overall slightly more likely to be reviewed than non-collective action events.

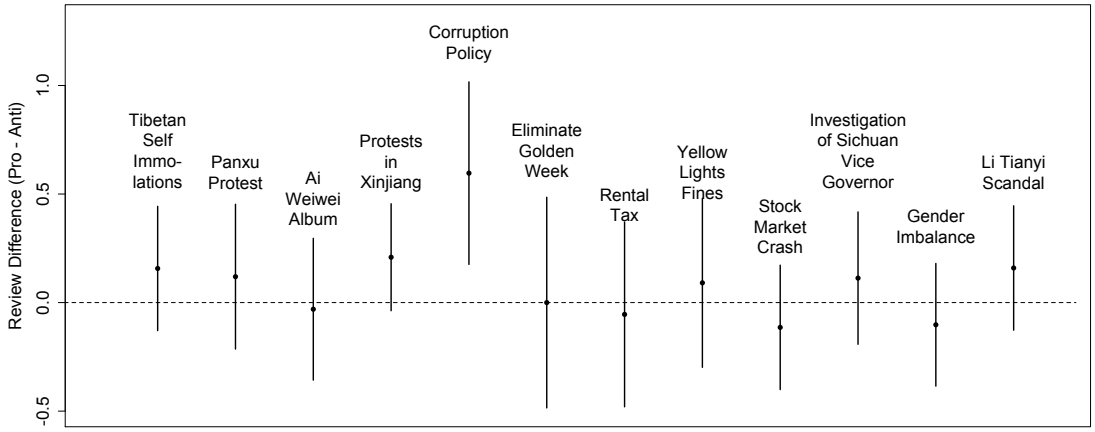

Figure 6: Causal Effect on Review of Posts For or Against the Government: Automated review picks up slighly more posts that are for the government as compared to posts that are against the government.

nonzero relationship here, it is that submissions in favor of the government are reviewed more often than those against the government! Indeed, 9 of 12 point estimates are above zero, and two even have their entire confidence interval above zero. This seems like more of a mystery: government social media sites are slightly *more* likely to delay publication of submissions that favor the government, its leaders, or their policies. Private sites don't use automated review much at all. Why is this? We found that the answer again is the highly inexact keyword algorithms used to conduct the automated review.

To understand this better, we reverse engineer the Chinese keyword algorithms in order to discover the keywords that distinguish submissions reviewed from those not reviewed. Because the number of unique words written overwhelms the number of pub-

lished posts, we cannot find these keywords uniquely. However, we identify words highly associated with review using a "term frequency, inverse document frequency" algorithm [29, 30]. That is, we take the frequency of each word within the review posts and divide this number by the number of non-reviewed documents in which that same word appears. Thus for every word we have a measure of its frequency in review posts, relative to posts that were not reviewed. Words with high values on these measures are likely to be used within the automated review process.

Table 2 gives the top keywords (and keyphrases) we estimate were used to select posts we wrote into automated review. We can see that the words associated with review could plausibly detect collective action and relate to the government and its actions, but are also just as likely to appear in pro-government posts as in anti-government posts. For example, more pro- than anti-government posts are reviewed in the Corruption Policy topic in Figure 4. This appears to be because the reviewed pro-government posts used the word corruption (腐败) more frequently than anti-government posts. However, corruption was used in the context of praising how the new policy would strengthen anti-corruption (反腐败) efforts. Not only is automated review only conducted by a subset of websites and largely ineffective at detecting posts related to collective action events, but it also can backfire by delaying the publication of pro-government material.

It turns out that we can also offer a test of the veracity of these keywords. In the context of setting up our own web site, we unearthed a list of keywords for review that a software provider offered to their clients running social media web sites. The list is dated to April 2013, and all of the keywords we found related to events taking place prior to April 2013 were on this list. The exceptions were from events that occurred after April 2013.

It thus appears that the workers in government-controlled web sites are so risk adverse that they have marshaled a highly error prone methodology to try to protect themselves. They apparently know not to take this automated review methodology very seriously as, whether it is used or not, the manual process of reading individual posts is still used widely since our results show that automated reivew does not affect the causal effect of collective

Chinese	English
群众	masses
政府	government
事件	incident
恐怖	terror
新疆	Xinjiang
中国	China
上街	go on the streets
李天一	Li Tianyi
法律	law
达赖	Dalai Lama
游行	demonstration
香港	Hong Kong
行贿	to bribe
腐败	corruption

Table 2: Top keywords distinguishing posts held vs not held for review. Words within this list match keyword lists provided by the software provider.

action events on censorship decisions.

4 Edge Cases

We now attempt to define the outer boundaries of the theory of collective action potential by choosing cases close to, but outside, the theory and look for no effect, and one inside (criticism of the top leaders) but extreme.

Internet-Only and External-Only Collective Action The first case is an event that had collective action taking place but only on the Internet. At the end of May, 2013, the principal of Hainan Wanning City No. 2 Elementary School was being investigated for taking six elementary school girls to a hotel. Ye Haiyan, a women's rights advocate went to the elementary school and protested with a sign in her hand that read "Principal: get a hotel room with me, let the elementary students go! Contact Telephone: 12338 (Ye Haiyan)." Ye's protest went viral and her sign became an online meme, where netizens would take and share photos of themselves, holding a sign saying the same thing with their own phone numbers or often with China's 911 equivalent (110) as the contact phone

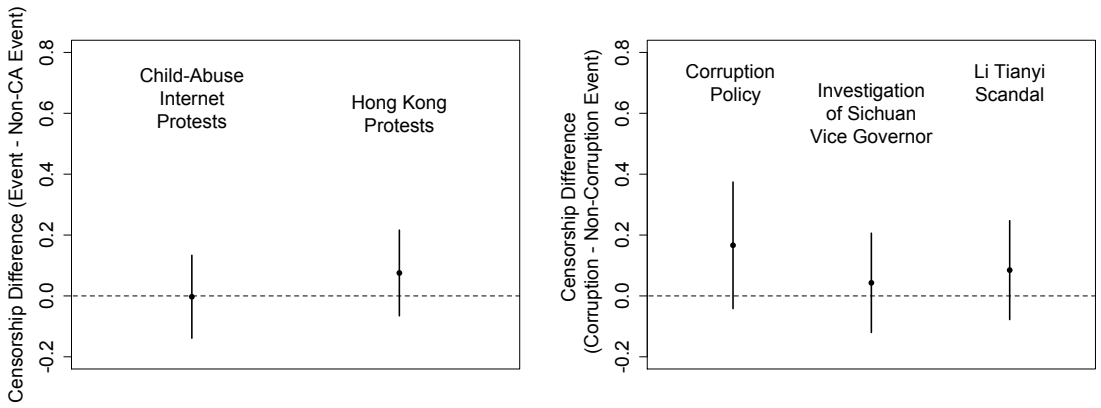

Figure 7: Testing Edge Cases for the Causal Effect of Collective Action Potential (left panel) and of Posts About Corruption (right panel)

number.[6]

The second event occurred on July 1, 2013, which was the 16th anniversary of the handover of sovereignty of Hong Kong from Britain to China. Every year on this day, thousands take to the streets of Hong Kong in protest, but typically with little or no such protest on the mainland. In 2013, between 30,000 people (according to the police) and 430,000 people (according to the organizers) took to the streets to call for true democracy and Chief Executive CY Leung's resignation.[7]

Neither of these "edge case" examples meet the definition of collective action events given in Section 2, but they are obviously close. We ran our experimental design for these events too, and give the results in the left panel of Figure 7. In both cases, the overall causal effect is near zero, with confidence intervals that overlap zero. There is a hint of a possibly positive effect only for posts reviewed about Hong Kong protests, but in the context of the natural variability of Figures 2 and 3 is not obviously different from zero.

Corruption and Wrong-Doing among Senior Leaders Next, we study the effects of writing about corruption and wrong-doing among senior leaders in the government, Party, and military on censorship. Nothing in the theory of collective action potential supports this effect but, because corruption so directly implicates leaders who could control censoring, considerable suspicion exists in the literature that posts about corruption are censored

[6]For examples see [31].

[7]For news coverage of the protests, see [32, 33, 34, 35].

[12, 20, 14]. We can even point to the odd result regarding this topic that posts supporting the government's efforts to deal with corruption are more censored than posts opposed to the government's efforts to deal with corruption (see Figure 6).

We selected three corruption-related topics for the analysis. The first relates to a new corruption policy that imposes criminal charges against bribes exceeding 10,000 Chinese yuan. The second topic relates to the investigation of Guo Yongxiang, a member of the Sichuan Province Central Committee and a Vice-Governor of Sichuan for serious breaches in discipline. The final topic relates to the naming of Li Tianyi, the son of a well-known People's Liberation Army performer Li Shuangjiang, for participating in a gang-rape. The Li Tianyi case led to speculations of corruption that Li's father's ties to the People's Liberation Army would allow Li to avoid punishment commiserate with his crimes. The results for an analysis of three corruption events appear in the right panel of Figure 7, all of which clearly show no effect, thus again supporting the theory of collective action potential. Similarly supportive is the fact that posts in these topics name specific Chinese government and CCP leaders, both at central and local levels of government (see the online supplement).

Top Leaders and Highly Sensitive Issues Finally, we study the question of the censorship of discussions about top Chinese leaders, arguments for deep political reform, and discussion of highly sensitive or salient issues. We study these topics with observational methods.

To study more directly whether Chinese censors allow direct criticism of top leaders, we begin by finding a social media volume burst about Chinese President Xi Jinping that (1) by our specific definitions does not have collective action potential, (2) includes posts that cover meaningful and important topics, and (3) is about a topic that could generate highly critical posts about the leader. We found the following volume burst that met these conditions.

On December 28, 2013, President Xi Jinping visited a Feng Qing Steamed Bun Shop in Beijing (Feng Qing is a chain restaurant), and ate steamed buns "just like the rest of us." He waited in line, he paid 21 CNY for steamed pork and onion buns along with a

side of stir-fried liver, and he brought his own tray to a table. Xi's visit unleashed a storm of traditional media coverage and a large volume burst on social media. Although Xi's visit to the bun shop sounds like an innocuous event, online discussions related the visit to important and high profile issues such as Xi's China Dream, corruption of government officials, rising real estate prices, the plight of China's elderly and impoverished, as well as propaganda, censorship, the absence of elections, and multi-party competition. However, this event is not connected to any on-going collective action events.

During this volume burst, we collected 82,280 social media posts related to this event before any posts were censored, and then checked each one from a network of computers around the world which were eventually censored. Finally, we applied the Hopkins-King algorithm [36] (using a training set of 592 hand coded posts) to determine the proportion of censored posts that were critical vs. supportive posts, and applied the Bayesian algorithm derived in [2] to invert this. We found, consistent with the collective action potential hypothesis, that posts critical of President Xi were censored just about as much as those which were supportive. Among posts that are critical of Xi and his actions, 18% (with a 95% confidence interval of 13–22%) are censored. Among the posts supportive of Xi the percent censored is 14% (with a 95% confidence interval of 8–22%). (The proportion of posts censored among posts that simply describe the event is 21%, with a confidence interval of 18-24%).

The supplementary materials include the text of examples of uncensored posts that are highly critical of President Xi and which use this event to discuss important issues. These posts involve many vivid personal attacks on Xi and his policies. In our experience, these posts are not surprising or unusual.

Next, we look for uncensored discussion of deep political reform. In August of 2013, three commentaries were published in *People's Daily* condemning constitutionalism, describing constitutionalism as incompatible with socialism and doomed to fail in China. These commentaries sparked a social media volume burst with intensive online discussions about whether China should adopt American style constitutionalism and multi-party competition. In the days following these commentaries, we collected a random sample of

9,850 blog posts related to political reform. While there are posts that toe the party line and criticize constitutionalism, there are also many uncensored posts advocating for the adoption of multi-party competition, describing reform as the only way to empower the Chinese people and to rein in corruption. We include several examples in our supplementary materials.

Finally, we sought and identified social media volume bursts related to three highly salient and politically sensitive issues about real world events that did not have collective action potential. These are discussions related to Tibet, Uyghurs, and Ai Weiwei.

First is the case of a volume burst in Tibet: In early August 2013, a post by a woman who claimed to have spurred her true love in order to marry a man who lived within view of Lhasa's Potala Palace went viral. As expected, censorship of posts in this burst is low at 12%.

Second is a volume burst related to Xinjiang and Uyghurs, which occurred in March 2013, where a post poking fun at a government entity with an exceptionally long, 54 character name (新疆维吾尔自治区乌鲁 木齐国家高新技术产业开发区社会管理综合治理委员会学校及周边治安综合治 理工作领导小组办公室) went viral. The government entity is located in Xinjiang, and the name can be roughly translated as the "Public Security and Management Office of the Working Small Group of the Holistic Social Management Committee's School and Surrounding Areas of Xinjiang Uyghur Autonomous Region Urumqi's Chinese High Tech Development Zone." This post was the butt of jokes and satire related to Chinese government bureaucracy, but completely unrelated to any on-going collective action event. As expected, censorship of this volume burst was low, only 10%.

Finally, we identified a volume burst related to artists, including Ai Weiwei along with Matisse, Picasso, Andy Warhol and others, and their cats. Censorship of this burst was also low, at 6%. Our definition of collective action potential includes real world events related to those who have catalyzed or organized collective action in the past. This volume burst relates to Ai Weiwei but falls outside the definition because the burst is not related to a real world event, nor is it solely related to Ai Weiwei.

5 Concluding Remarks

We offer the first large scale randomized experimental analysis of censorship in China, along with participant observation of how censorship is conducted. We use these designs to conduct a rigorous test of the theory of collective action potential, and to further uncover and resolve academic conflicts about crucial aspects of the Chinese censorship program. With them we are able to subject to empirical estimation what had previously been left to statistical assumption. We are also able to study the large program whereby enormous numbers of social media submissions are put into limbo before being considered for possible publication or censorship. Whereas censorship is a publish-first-censor-later process, automated review involves a review-first-maybe-publish-later process.

Our flexible research designs also enabled us to study edge cases, just beyond the reigning theory of collective action potential, so that we can define the boundaries of where it applies. This includes the effects of highly salient and sensitive topics about events without collective action potential; posts about corruption; posts that name Chinese leaders specifically; and collective action events that are solely on the Internet or about collective action on the ground outside the Chinese mainland — none of which are predicted by the theory of collective action potential to be censored more than others, and which our data clearly shows is not censored more than other non-collective action topics. We also show that academic controversies over confusing interpretations of which keywords are being censored in automated review is resolved once we realize that the Chinese government is surprisingly flexible over what methods and technology each social media site can use, even while imposing uniformity of results by requiring post hoc censoring by human coders.

Future researchers should consider comparing these results on censorship in social media with censorship in traditional media and other ways the Chinese government impedes the free flow of information.

References

[1] Freedom House, "Freedom of the press, 2012." www.freedomhouse.org, 2012.

[2] G. King, J. Pan, and M. E. Roberts, "How censorship in china allows government criticism but silences collective expression," *American Political Science Review*, vol. 107, pp. 1–18, 2013. http://j.mp/LdVXqN.

[3] S. Shirk, *China: Fragile Superpower: How China's Internal Politics Could Derail Its Peaceful Rise*. New York: Oxford University Press, 2007.

[4] S. L. Shirk, *Changing Media, Changing China*. New York: Oxford University Press, 2011.

[5] M. Whyte, *Myth of the Social Volcano: Perceptions of Inequality and Distributive Injustice in Contemporary China*. Stanford, CA: Stanford University Press, 2010.

[6] L. Zhang, A. Nathan, P. Link, and O. Schell, *The Tiananmen Papers*. New York: Public Affairs, 2002.

[7] A. Esarey and Q. Xiao, "Political expression in the chinese blogosphere: Below the radar," *Asian Survey*, vol. 48, no. 5, pp. 752–772, 2008.

[8] R. MacKinnon, *Consent of the Networked: The Worldwide Struggle For Internet Freedom*. New York: Basic Books, 2012.

[9] P. Marolt, "Grassroots agency in a civil sphere? rethinking internet control in china," in *Online Society in China: Creating, Celebrating, and Instrumentalising the Online Carnival* (D. Herold and P. Marolt, eds.), pp. 53–68, New York: Routledge, 2011.

[10] M. Dimitrov, "The resilient authoritarians," *Current History*, vol. 107, no. 705, pp. 24–29, 2008.

[11] X. Chen, *Social Protest and Contentious Authoritarianism in China*. New York: Cambridge University Press, 2012.

[12] D. Bamman, B. O'Connor, and N. Smith, "Censorship and deletion practices in chinese social media," *First Monday*, vol. 17, no. 3-5, 2012.

[13] T. Zhu, D. Phipps, A. Pridgen, J. Crandall, and D. Wallach, "The velocity of censorship: High-fidelity detection of microblog post deletions," in *22nd USENIX Security Symposium*, 2013.

[14] R. MacKinnon, "China's censorship 2.0: How companies censor bloggers," *First Monday*, vol. 14, no. 2, 2009.

[15] K. Hong, "China's twitter-like sina weibo service now has over 50 million active users per day," August 13 2013. http://tnw.co/1fdNFPS.

[16] S. Millward, "Tencent weibo, the 'other weibo' that noboday cares about, reaches 540 million users," January 22 2013. http://bit.ly/1byjSNW.

[17] K.-w. Fu and M. Chau, "Reality check for the chinese microblog space: A random sampling approach," *PLoS One*, vol. 8, no. 3, 2013.

[18] K. Imai, G. King, and E. Stuart, "Misunderstandings among experimentalists and observationalists about causal inference," *Journal of the Royal Statistical Society, Series A*, vol. 171, part 2, pp. 481–502, 2008. http://gking.harvard.edu/files/abs/matchse-abs.shtml.

[19] J. Ng, "Weibo keyword un-blocking is not a victory against censorship," June 21 2013. http://bit.ly/1kfqNBC.

[20] J. Crandall, M. Crete-Nishihata, J. Knockel, S. McKune, A. Senft, D. Tseng, and G. Wiseman, "Chat program censorship and surveillance in china: Tracking tomskype and sina uc," *First Money*, vol. 18, no. 7-1, 2013.

[21] J. Fallows, ""the connection has been reset"," *The Atlantic*, 1 March 2008. http://goo.gl/U56yfw.

[22] K. Imai, L. Keele, D. Tingley, and T. Yamamoto, "Unpacking the black box of causality: Learning about causal mechanisms from experimental and observational studies," *American Political Science Review*, vol. 105, no. 4, pp. 765–789, 2011.

[23] J. Pearl, "Direct and indirect effects," in *Proceedings of the seventeenth conference on uncertainty in artificial intelligence*, pp. 411–420, 2001.

[24] O. Blanchard and A. Shleifer, "Federalism with and without political centralization: China versus russia," tech. rep., National Bureau of Economic Research, 2000.

[25] S. Heilmann and E. Perry, *Mao's Invisible Hand: The Political Foundations of Adaptive Governance in China*. Cambridge, MA: Harvard University Asia Center, 2011.

[26] Y. Qian and G. Roland, "Federalism and the soft budget constraint," *American economic review*, pp. 1143–1162, 1998.

[27] Y. Qian and B. R. Weingast, "Federalism as a commitment to perserving market incentives," *The Journal of Economic Perspectives*, vol. 11, no. 4, pp. 83–92, 1997.

[28] G. Egorov and K. Sonin, "Dictators and their viziers: Endogenizing the loyalty-competence trade-off," *Journal of European Economic Association*, pp. 903–930, 2011.

[29] G. Salton, *Automatic Text Processing: the transformation, analsis, and retrieval of information by computer*. Reading, Mass.: Addison-Wesley, 1988.

[30] D. Kelleher and S. Luz, "Automatic hypertext keyphrase detection," in *International Joint Conference on Artificial Intelligence*, vol. 19, p. 1608, Lawrence Erlbaum Associates, 2005.

[31] P. Barefoot, "Principal, get a room with me, spare the schoolchildren!," May 31 2013. http://j.mp/19yuv7E.

[32] Aljazeera, "Democracy push as hong kong marks handover," July 1 2013. http://j.mp/145Jvpp.

[33] S. Lee and K. Wong, "Hong kong protests to underscore leung's record-low appeal," June 28 2013. http://j.mp/13r3v7v.

[34] J. Ngo, "July 1 protest is hong kong's taste of democracy," June 30 2013. http://j.mp/15PcwBt.

[35] C. Yung, "Annual hong kong protest to focus ire on leader," June 28 2013. http://j.mp/13FJB3w.

[36] D. Hopkins and G. King, "A method of automated nonparametric content analysis for social science," *American Journal of Political Science*, vol. 54, pp. 229–247, January 2010. http://gking.harvard.edu/files/abs/words-abs.shtml.